A Multi-Document Multi-Lingual Automatic Summarization System

Mohamad Ali Honarpisheh, Gholamreza Ghassem-Sani, Ghassem Mirroshandel

Sharif University of Technology,

Department of Computer Engineering, Tehran, Iran,
Honarpisheh@ce.sharif.edu, Sani@sharif.ir, Mirroshandel@ce.sharif.edu

Abstract

Abstract. In this paper, a new multi-document multi-lingual text summarization technique, based on singular value decomposition and hierarchical clustering, is proposed. The proposed approach relies on only two resources for any language: a word segmentation system and a dictionary of words along with their document frequencies. The summarizer initially takes a collection of related documents, and transforms them into a matrix; it then applies singular value decomposition to the resulted matrix. After using a binary hierarchical clustering algorithm, the most important sentences of the most important clusters form the summary. The appropriate place of each chosen sentence is determined by a novel technique. The system has been successfully tested on summarizing several Persian document collections.

1 Introduction

With the advent of the Internet, different newspapers and news agencies regularly upload their news in their sites. This let users to access different viewpoints and quotes about a single event. At the same time of this explosive growth of the amount of textual information, the need of people for quick access to information has dramatically increased. The solution proposed for dealing with this huge amount of information is using Text Summarizers. Several systems have been developed with respect to this solution (McKeown et. al., 2002; Radev et. al., 2001).

Generally in the process of multi-document text summarization, a collection of input documents about a particular subject is received from the user and a coherent summary without redundant information is generated. However, several challenges exist in this process the most important of which are removing redundant information from the input sentences and ordering them properly in the output summary. In a new approach to multi-document summarization proposed in this paper, Singular Value Decomposition (SVD) is used to find the most important dimensions and also to remove noisy ones. This process makes clustering of similar sentences easer. In order to determine the level of importance of different clusters, the generated singular values and singular vector of the SVD have been used in a fashion similar to that (Steinberger and., Ježek, 2004). To evaluate generated summaries the SVD-based method proposed in the same paper is used.

2 Text Summarization approaches

There are different features with which we can classify text summarization systems. In (Sparck-Jones, 1999) these features are divided according to the input, purpose, and output of system. With respect to this categorization, our proposed system is a general multi-document multi-lingual text summarizer which generates extracts a summary from the input documents.

Different approaches to text summarization can be categorized in different ways according to various features of text summarization systems. With

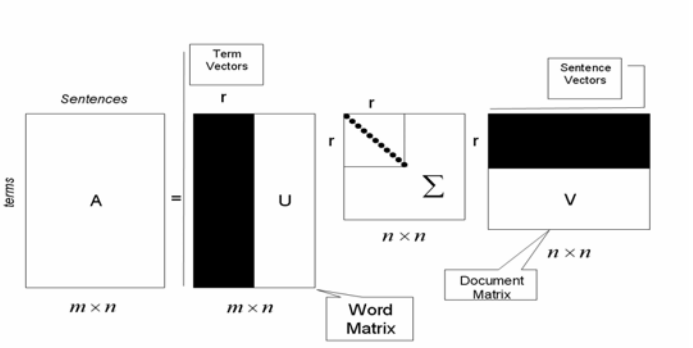

Fig. 1. Singular Value Decomposition

respect to the output of the system, there are two categories of extracting and abstracting methods. Extraction-based summarization methods are also divided into three classes.

The first group of Extraction-based methods is statistical. These methods statistically assign significance score to different textual units. The very first developed summarization methods were of this category (Edmundson and Wyllys, 1961). Scoring policy in these systems was based on different features, such as term frequency and place of the sentences. Vector Space Models (Salton et. al., 1994), compression of sentences with Automatic Translation approaches (Knight and Marcu, 2000), Hidden Markov Model (Jing and McKeown, 2000), Topic Signatures based methods (Lin and Hovy, 2000, Lacatusu et al., 2006) are among the most popular techniques that have been used in the summarization systems of this category.

The second groups of extraction-based methods, shallow understanding approaches use some information about words or textual units and their dependencies to improve the performance of extraction. The understanding can be induced using dependencies between words (Barzilay and Elhadad, 1997), rhetorical relations (Paice and Johns, 1993), events (Filatova and Hatzivassiloglou, 2004). In all of these methods, the most focused dependencies are used as a measure for saliency of each textual unit.

The third group, knowledge-based approaches, uses particular domain knowledge in discriminating the important parts of input documents. This knowledge is usually taking some assumptions about the working domain. Centrifuger (Elhadad et al., 2005) is a good example of systems in this category, which operates in medical domains.

The new approach proposed in this paper uses SVD and hierarchical clustering methods. It can therefore be categorized in the statistical based methods.

3 SVD based methods

Methods that use SVD to find salient information in the input document are a member of Vector Space Models. In such models, each textual unit is represented by a vector. Each component of this vector is filled with a value which represents both the local and global importance of each word.

The idea of using SVD in summarization was first introduced in (Gong and Liu, 2001). In this model, the input document is transformed into an $m \times n$ sparse matrix, *A*, where *m* is the number of words and *n* is the number of the sentences of input document.

The SVD of this $m \times n$ matrix, with the assumption $m \succ n$, is defined as follows:

$$A = U \Sigma V , \qquad (\mathbf{1})$$

where *U=[u_{ij}]* is an $m \times n$ column-orthogonal matrix with columns named as the left singular vectors, $\Sigma = diag(\sigma_1, \sigma_2, ..., \sigma_n)$ is an $n \times n$ diagonal matrix with non-negative diagonal elements in descending order, and *V=[v_{ij}]* is an $n \times n$ row-orthogonal matrix with rows named as the right singular vectors (figure 1 demonstrate application of SVD to *A*). The number of non-zero elements in Σ is equal to the rank, *r*, of matrix *A*.

There are two viewpoints about the result of performing SVD on sentence by the word matrix of document (Gong and Liu, 2001). From transformation viewpoint, SVD for each sentence reduces the dimensions from *m* to *r*. The salience degree of the reduced dimension decreases from the first to the *r*th dimension. From the semantic viewpoint, SVD derives the hidden latent structure of the input document. This structure is represented in *r* linearly independent base vectors (i.e. concepts). SVD can capture and model the interrelations between concepts and hidden topics, which can be used to cluster semantically related words and sentences.

In SVD-summarization method, for each salient concept (singular vector) of matrix V, the sentence with the highest value in that dimension is chosen. This technique helps us to choose the sentences of the summary that best represent the most important concepts. For example, the most important sentence of a document in this summarization method is the sentence that has the highest value in the first row of *V*.

However, this method faces two significant problems. First, the number of dimensions should be less than or at most equal to the number of discovered. Second, in this method just individual concept is used to determine their saliency, not their combinations. This strategy obviously works poor when a sentence that is not the most important of any dimension alone may contain concept that in combination make it important enough.

These problems led to the introduction of a new summarization method (Steinberger and Ježek, 2004). This new approach uses summation of the weighted components of singular vectors instead of each individual concept alone. The weight of each vector's component of each sentence is its corresponding singular value. The reason for such weighting is to increase the effect of more important singular vectors. Formally, the degree of salience of each sentence can be computed by using the following formula:

$$s_k = \sqrt{\sum_{i=1}^{r} v_{k,i}^2 . \sigma_i^2 .} \qquad (2)$$

where s_k is the salience degree of *k*th sentence in the modified latent vector space, and *r* is the number of important dimensions of the new space. Corresponding value of each *r* dimensions is greater than half of the first singular value.

Both of above strategies for text summarization were proposed for single document summarization only. These approaches do not utilize the clustering power of SVD in discovering sentences with close meanings. On the other hand, their pure reliance on SVD, which does not depend on the characteristics of any language, makes them appropriate to be applied to any language.

4 Multi-Document SVD based Summarization

In this paper a new version of the SVD based summarization is introduced. After transforming all documents into sentence by word matrix, SVD is applied to the resultant matrix. To remove redundant sentences from the summary, a hierarchical bottom-up clustering algorithm is applied to the *r* most important extracted concepts of the input sentences. After extracting the clusters, the saliency score of each cluster is determined, and the most important clusters are selected. At the same time, using a simple ordering method, the appropriate place of each sentence in the sorted collection is determined. In the following sections, each of these processes is described in more details.

Matrix Construction and Application of SVD

Given a collection of input documents that are related to a particular subject, the system decomposes the documents into sentences and their corresponding words. In addition, it computes the total number of occurrences of each word in all documents as the Term Frequency (TF).

The developed system works on Persian language. It is assumed that words are separated by spaces. However, some words in Persian are compound words .However this did not cause any problem for the developed system; because the most meaningful part of such words is usually less common than others and thus have an Inverse Document Frequency (IDF) that is higher than that of more common less meaningful parts. IDF represents the amount of meaning stored in each word. It can be used to reduce the impact of less important constituents such as stop words, which usually have a high TF but contains little meaning.The formula for calculating IDF is as follows:

$$IDF(term) = \log(\frac{NUMDOC}{NUMDOC(term)}) , \qquad (3)$$

where *NUMDOC* represents the total number of document processed to create IDF dictionary and *NUMDOC(term)* is the number of documents in which the *term* appeared.

After decomposition, the input sentences along with their corresponding words are arranged as a matrix. Two different weighting schemes have been applied to each element of this matrix: 1) a constant value and, 2) each word's associated TF*IDF (Salton and Buckley, 1988).

After constructing the Sentence by word matrix, SVD is applied to the resultant matrix. Applying SVD removes the effect of unimportant words and highlights more salient dimensions. It also reduces the number of dimensions for each sentence, resulting in an easer clustering process. This improves the performance of sentence clustering by making it faster and less sensible to unimportant differences between sentences.

Clustering

To cluster reduced dimension sentences, a binary hierarchical agglomerative clustering algorithm with average group linkage is used. In this algorithm, at first, each sentence is considered as a cluster. At each step, two closest clusters are combined into a single cluster. The dimension of this new cluster is the average dimensions of the two combining ones. These steps are repeated until we have only one cluster. So the result of this algorithm is a binary tree.

The question that needs to be answered at this step is "how clusters containing similar sentences can be extracted from this binary tree?" Two properties are required to propose a sub-tree as a possible cluster of similar sentences:

1. The number of existing sentences at the current node (cluster) should be less than or equal to the total number of input documents; because it is assumed that there is not much redundant information in each document. This assumption is valid with respect to the news documents in which there might be only little redundancy.
2. The distance between two children of the current cluster should be less than or equal to the distance between current cluster and its sibling node. This condition has been found empirically.

Using these two heuristics, similar clusters are extracted from the binary tree.

Finding Salient units

To select important clusters from the set of extracted clusters, different clusters are scored based on two diffrent methods. In the first method, the average of TF*IDF of different words in the each sentence in the current cluster are used. The second approach was the latest SVD-based approach which was proposed by (Steinberger and Ježek, 2004) and was described in the section 3. In the latter, score of each cluster is found using the following formula:

$$score(cluster) = \frac{\sum_{s \in cluster} score(s)}{|cluster|}, \qquad \textbf{(4)}$$

where $|cluster|$ represent the number of sentences in the current cluster.

Selecting and ordering sentences:

In this step, the final representation of the summary will be generated. To this end, from each important cluster, a sentence should be selected. At the same time, the proper order of selected sentences should be determined. To find this order, a Representative Document (RD) is selected. The RD is the document which includes most sentences of the most important clusters. After selecting RD the following steps are performed:

1. Starting from the most important cluster, while the number of processed words of summary sentences does not exceed form the specified number:
   a. If no sentence from the current cluster was not added to the summary:
      i. If there is a sentence from the RD in this cluster, choose this sentence;
      ii. Otherwise find the most important sentence, the current cluster: To find out the place of the selected sentence in the summary, a search is performed for clusters that contain both sentences from RD and neighbors of the selected sentence. The place of the sentences from RD that co- clustered with neighbors of the selected sentence is chosen as the selected sentence boundary.
      iii. If any place has been found for the selected sentence, add it to summary in the specified location, and mark that cluster as having a sentence in the summary.

2. If it remains any unplaced sentence which should be presented in the summary, go to step 1 with the remaining number of words.

5 Experiments

5.1 Testing Collection

The proposed summarizer is originally developed for the Persian language. In Persian like many other languages there is not a standard test collection, to evaluate the summarizers. To overcome the lake of a test collection in Persian, an unsupervised approach of evaluating summaries is selected (i.e. SVD-based evaluation method proposed in (Steinberger and Ježek, 2004)). In addition to an evaluator, a collection of documents was also required. For this purpose different texts related to a single event were collected. The properties of these collections are presented in table 1

5.2 Results and Discussions:

To find out which term weighting and distance measure causes the highest increase in the SVD-Scores, various combinations of these approaches has been used in the summarization approach. To find the distance between clusters, Euclidian, Hamming, and Chebyshev distances and to determine the saliency of different clusters, TFIDF and SVD-based methods were used. The gained SVD-Based score using different configurations are represented in table 2

As it can be seen in table 2, TFIDF-based methods score higher than SVD-based methods. Also, the most promising distance was the hamming distance. It can also be seen that the performance decreases substantially when instead of a constant value *tf*idf* scores were used. It was observed that using various distance measure the SVD-Score of each collection would be different. The SVD-Scores are in favor of using the boosting methods for classification of sentences with different distance measure for each classifier. Comparing these results with the ones proposed in (Steinberger and Ježek, 2004), a significant decrease in evaluated SVD-Based scores is observed. One of the reasons for this phenomenon is that the distinct words appearing in multi-documents are more extensive than the words appear in a single document.

6 Conclusions

This paper presents a new SVD-based multi-lingual multi-document summarization method using agglomerative clustering. In this method, first, using SVD, the most important concepts representing sentences are extracted into a word by a sentence matrix. Then, similar sentences are clustered using a new heuristic method. Finally, impor-

Average number of distinct words in documents	1474
Average number of sentences in documents	32
Average number of sentences in subjects	643
Maximum Number of distinct words in subjects	2189
Minimum Number of distinct words in subjects	485
Number of subjects	14

Table 1. Testing Collections –Details

	Euclidian			Hamming			Chebyshev		
	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min
TFIDF	0.450	0.572	0.286	**0.518**	0.596	0.343	0.475	0.605	0.264
SVD-based	0.466	0.632	0.313	0.472	0.650	0.322	0.449	0.620	0.309

Table 2. Using a constant value for word-sentence matrix

	Euclidian			Hamming			Chebyshev		
	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min
TFIDF	0.364	0.549	0.109	0.269	0.512	0.113	0.406	0.512	0.283
SVD-based	0.309	0.134	0.563	0.319	0.499	0.112	0.367	0.518	0.235

Table 3. Using TF-IDF for each element of the matrix

in the summary are extracted from the resulting clusters. Different weighting schemes, distance metrics and scoring methods have been experimented. According to our experiments constant weighting scheme along with hamming distance is superior to other combinations. Since this method only needs determination of words and their inverse document frequency, it can be applied to any language providing these resources. We are now trying to improve the performance of the proposed algorithm. It seems that applying Principle Direction Partitioning (Blei, 2002) algorithm in the clustering phase and using Latent Dirichlet Allocation method (Boley, 1998) instead of the SVD based ones to model sentences can improve the score of the proposed method.

References

Barzilay, R. and Elhadad. M. 1997. “Using lexical chains for text summarization.” In Proceedings of the ACL/EACL'97 Summarization Workshop, Madrid, Spain.

Blei, D., Ng, A., and Jordan M., Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, January 2003. (A shorter version appeared in NIPS 2002).

Boley, D.L.: Principal Direction Divisive Partitioning. Data Mining and Knowledge Discovery, Vol. 2(4):325–344, Dec. 1998.

Edmundson, H.P. and Wyllys, R.E., Automatic abstracting and indexing - Survey and recommendations, Communications of the ACM, Vol. 4, (1961) 226-234

Elhadad, N., Kan, M.Y., Klavans, J., McKeown, K.: Customization in a unified framework for summarizing medical literature, Artificial Intelligence in Medicine Vol. 33(2): (2005) 179-198.

Filatova, E., Hatzivassiloglou, V.: Event-based Extractive summarization, In: Proceedings of ACL 2004 Workshop on Summarization, (2004) 104-111.

Gong, Y., Liu, X.: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. Proceedings of the 24th ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States (2001) 19-25

Jing, H., McKeown, K.: Cut and paste based text summarization. Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'00), (2000), Seattle-Washington

Knight, K., Marcu, D.: Statistics based summarization .step one: Sentence compression, Proceeding of the 17th National Conference of the American Association for Artificial Intelligence (2000)703-710.

Lacatusu, F., Hickl, A., Roberts, K., Shi, Y., Bensley, J., Rink, B., Wang, P., Taylor, L.: LCC’s GISTexter at DUC 2006: Multi-Strategy Multi-Document Summarization, Document Understanding Conference (2006)

Lin, C.Y., Hovy, E.: From single to multi-document summarization: A prototype system and its evaluation. Proceedings of the ACL, pages 457–464, 2002

McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J.L., Nenkova, A., Sable, C., Schiffman, B., Sigelman, S.: Tracking and summarizing news on a daily basis with Columbia's newsblaster. Proceedings of 2002 Human Language Technology Conference (HLT), San Diego, CA, 2002

Paice, C. D., Johns, A. P.: The identification of important concepts in highly structured technical papers, In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1993)

Radev, D. R., Blair-Goldensohn, S., Zhang, Z., Raghavan R.S.: Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization. Proceedings of 2001 Human Language Technology Conference (Demo Session), San Diego, CA, 2001

Salton, G., Allan, J., Buckley, C., Singhal, A.: Automatic analysis, theme generation, and summarization of machine readable texts, Science, Vol. 264(5164), (1994) 1421–1426

Salton, G. and Buckley, C.. Term weighting approaches in automatic text retrieval. Information Processing and Management, (1988), 24(5):513523

Sparck-Jones, K.: Automatic summarizing: factors and directions. Mani I, Maybury MT, editors. Advances in automatic text summarization. (1999) 10-12 [chapter 1]

Steinberger, J., Ježek, K. : Text Summarization and Singular Value Decomposition, Lecture Notes in Computer Science.Advances in Information Systems, Vol. 3261/2004, Springer-Verlag, Berlin Heidelberg New York (2004) 245-254